1. Music Classification by Genre
2. Project Summary: Music Classification by Genre
3. Music Classification by Genre: System Diagram
4. Introduction to Digital Signal Processing
5. Music Classification by Genre: Bandwidth
6. Music Classification by Genre: Frequency Cutoff
7. Music Classification by Genre: Frequency Smoothness
8. Music Classification by Genre: Beat Detection
9. Ideal Filters
10. Music Classification by Genre: High Pass Filter
11. Music Classification by Genre: Power Spectral Density
12. Music Classification by Genre: Total Power
13. Neural Networks
14. Music Classification by Genre: System Performance
15. Back propagation mathematics
16. Chris Hunter
17. Melodie Chu
18. Mitali Banerjee
19. Jordan Mayo

Music Classification by Genre

Project Summary: Music Classification by Genre

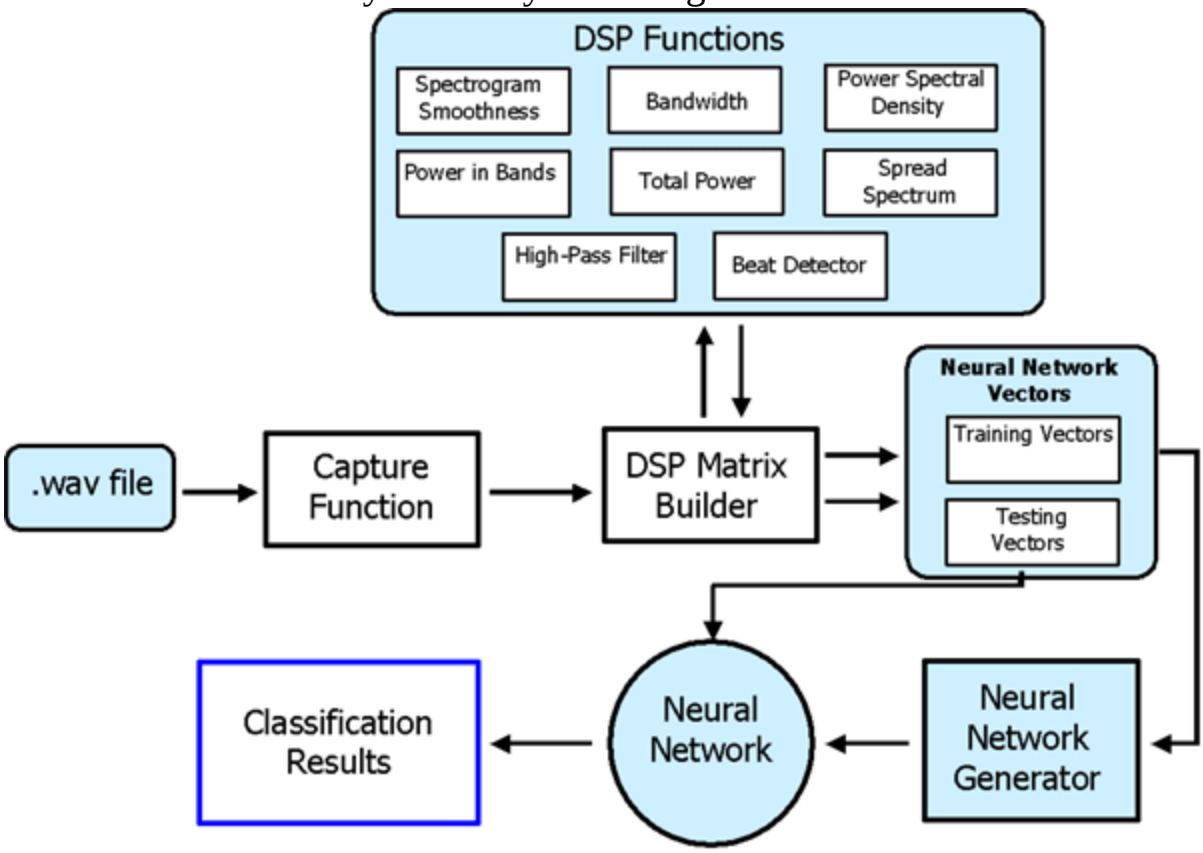
Widespread access to the Internet popularized digital music. People download large collections of music files sorted into directory structures by artist or genre. For example, a student at Rice University may want to search a library of files stored on the computer of a student in Bremen, Germany for classical music. Language differences and foreign preferences for file naming would make it difficult for the Rice student to determine music genres. A collection of filters to classify music based on DSP analysis tools would allow users to search a collection of files and extract only those that have certain chosen characteristics.

We designed a classification system that analyzes the contents of a .wav music file in order to sort it into specific categories: classical, jazz, country, rap, punk, and techno. In order to classify music samples, we examine characteristics in both the time and frequency domains:

- bandwidth
- beat(tempo) variability
- high pass filtering
- number of FFT coefficients above threshold
- power spectral density
- smoothness in frequency domain
- total power

Then a neural network classifies each song based on its similarity to other songs in various genres. Previous classification projects have directly analyzed song clips in neural networks. However, we take a slightly different approach by providing the neural network with the previously listed DSP characteristics that represent the song. This method proves 84% accurate, having most difficulty classifying techno music.

Music Classification by Genre: System Diagram
Music Classification by Genre System Diagram



Music Matcher, a collection of scripts and functions, takes a .wav file input, digitally processes it, and creates an output vector characteristic of the sample. A neural network is trained with 20 songs in each genre. Then it analyzes the new song vectors for patterns and predicts an output classification genre.

Music Matcher takes a .wav file, analyzes it, and outputs a music genre. Our system breaks up a .wav file into twenty .5 second windows. From here, the DSP functions are called for each of the twenty windows. Each one of these twenty windows is analyzed by seven DSP functions:

- Bandwidth
- Power Spectral Density
- Total Power (L-2 norm / L-infinity norm)

- Spectrogram Smoothness
- High Pass Filter
- Beat Detection
- Frequency Cutoff

The values returned from each of these functions is averaged over all twenty windows to give an average value for each song as well as a standard deviation, which tells us how these qualities change over time. That way, our DSP vector has some measure of how each of the functions changed with time.

First, the neural network is trained with 120 songs, 20 of each genre. After we train the neural network, we give it songs it has never seen, and the output of the system is the classification of genre that the neural network determines.

Introduction to Digital Signal Processing

Not only do we have analog signals --- signals that are real- or complex-valued functions of a continuous variable such as time or space --- we can define **digital** ones as well. Digital signals are **sequences**, functions defined only for the integers. We thus use the notation $s(n)$ to denote a discrete-time one-dimensional signal such as a digital music recording and $s(m,n)$ for a discrete-"time" two-dimensional signal like a photo taken with a digital camera. Sequences are fundamentally different than continuous-time signals. For example, continuity has no meaning for sequences.

Despite such fundamental differences, the theory underlying digital signal processing mirrors that for analog signals: Fourier transforms, linear filtering, and linear systems parallel what previous chapters described. These similarities make it easy to understand the definitions and why we need them, but the similarities should not be construed as "analog wannabes." We will discover that digital signal processing is **not** an approximation to analog processing. We must explicitly worry about the fidelity of converting analog signals into digital ones. The music stored on CDs, the speech sent over digital cellular telephones, and the video carried by digital television all evidence that analog signals can be accurately converted to digital ones and back again.

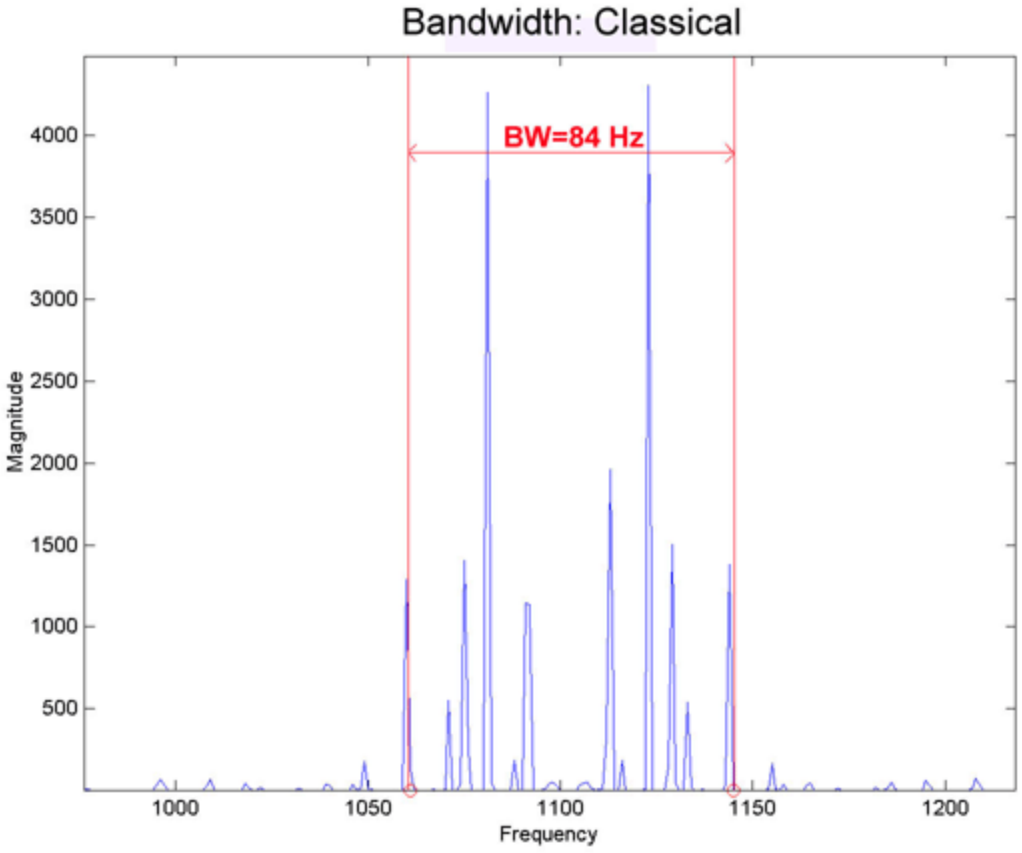
The key reason why digital signal processing systems have a technological advantage today is the **computer**: computations, like the Fourier transform, can be performed quickly enough to be calculated as the signal is produced, [footnote] and programmability means that the signal processing system can be easily changed. This flexibility has obvious appeal, and has been widely accepted in the marketplace. Programmability means that we can perform signal processing operations impossible with analog systems (circuits). We will also discover that digital systems enjoy an **algorithmic** advantage that contributes to rapid processing speeds: Computations can be restructured in non-obvious ways to speed the processing. This flexibility comes at a price, a consequence of how computers work. How do computers perform signal processing?

Taking a systems viewpoint for the moment, a system that produces its output as rapidly as the input arises is said to be a **real-time** system. All

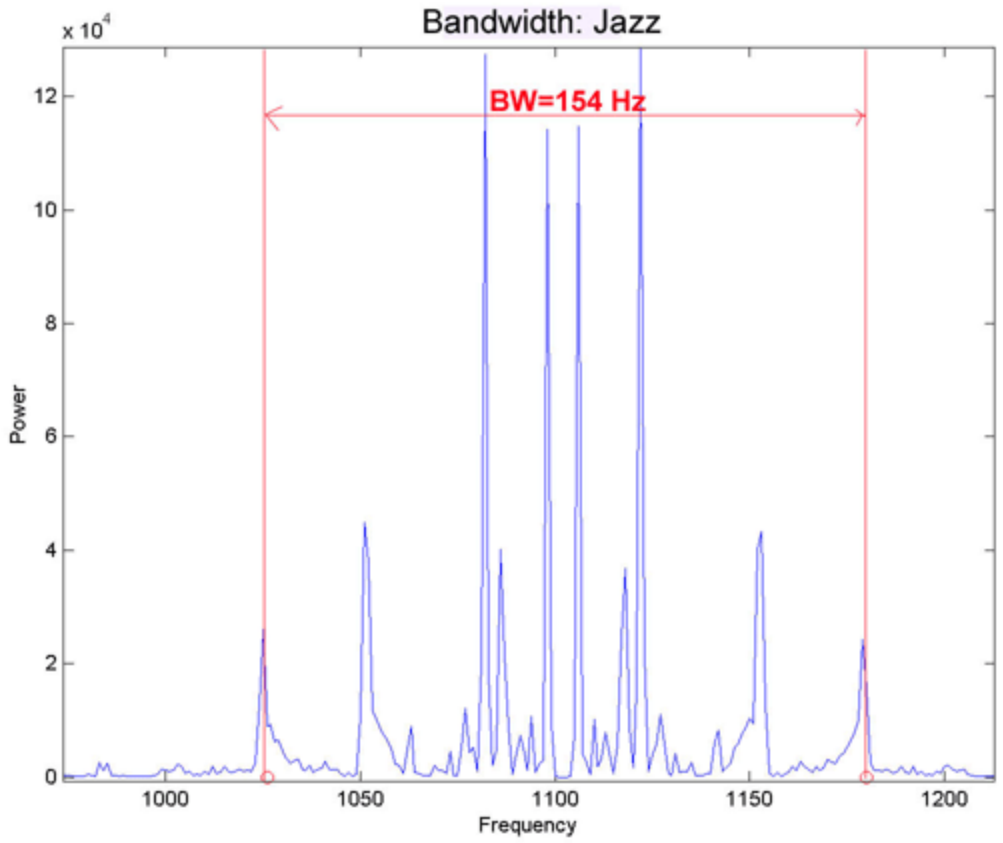
analog systems operate in real time; digital ones that depend on a computer to perform system computations may or may not work in real time. Clearly, we need real-time signal processing systems. Only recently have computers become fast enough to meet real-time requirements while performing non-trivial signal processing.

Music Classification by Genre: Bandwidth

Bandwidth refers to how spread-spectrum the signal is and what frequencies are present. If a signal is composed of many high frequencies, the bandwidth will be large. However, if the signal is composed of mostly low frequencies, the bandwidth will be small. After taking the shifted FFT of windows of the music vector, we find the last frequency component above a certain cutoff threshold, which is the bandwidth of the signal. Because classical music is composed of harmonic instruments, its bandwidth will be smaller and it will have fewer frequency components. However, hard music like punk or rap has lots of non-sinusoidal drumbeats, which will create more frequency components and their bandwidth will be larger.



Bandwidth for classical music.

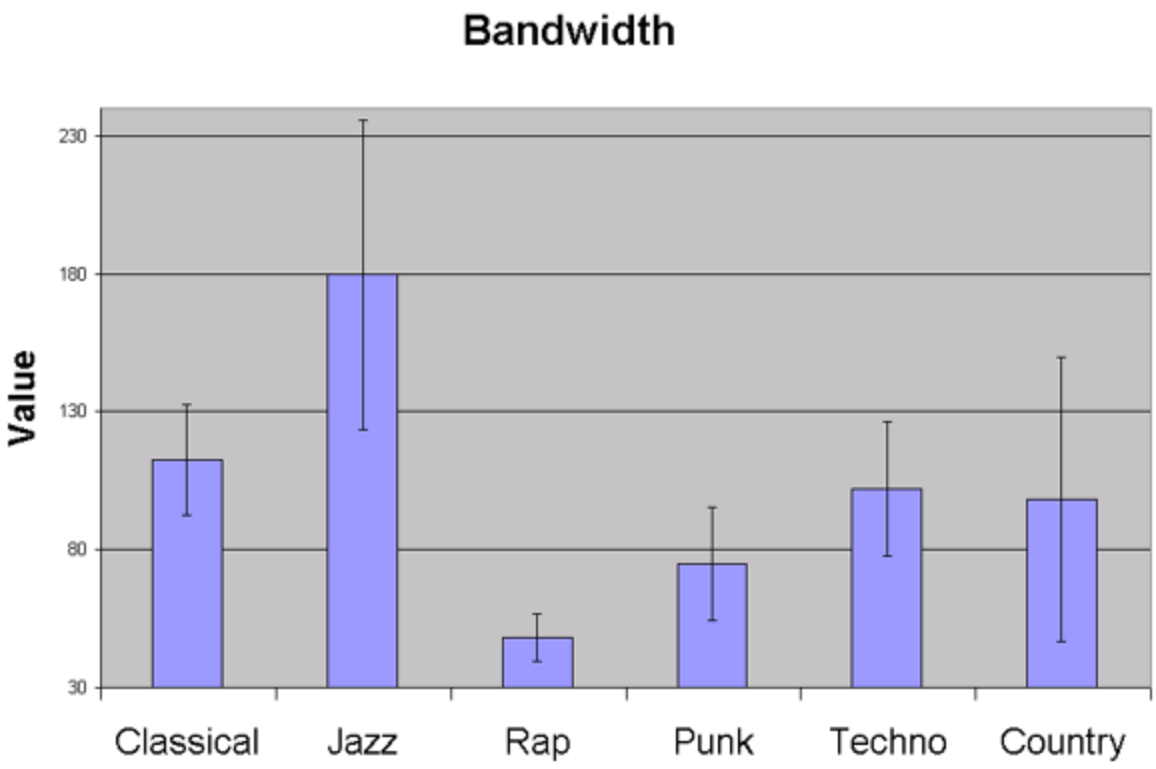


Bandwidth for jazz music.

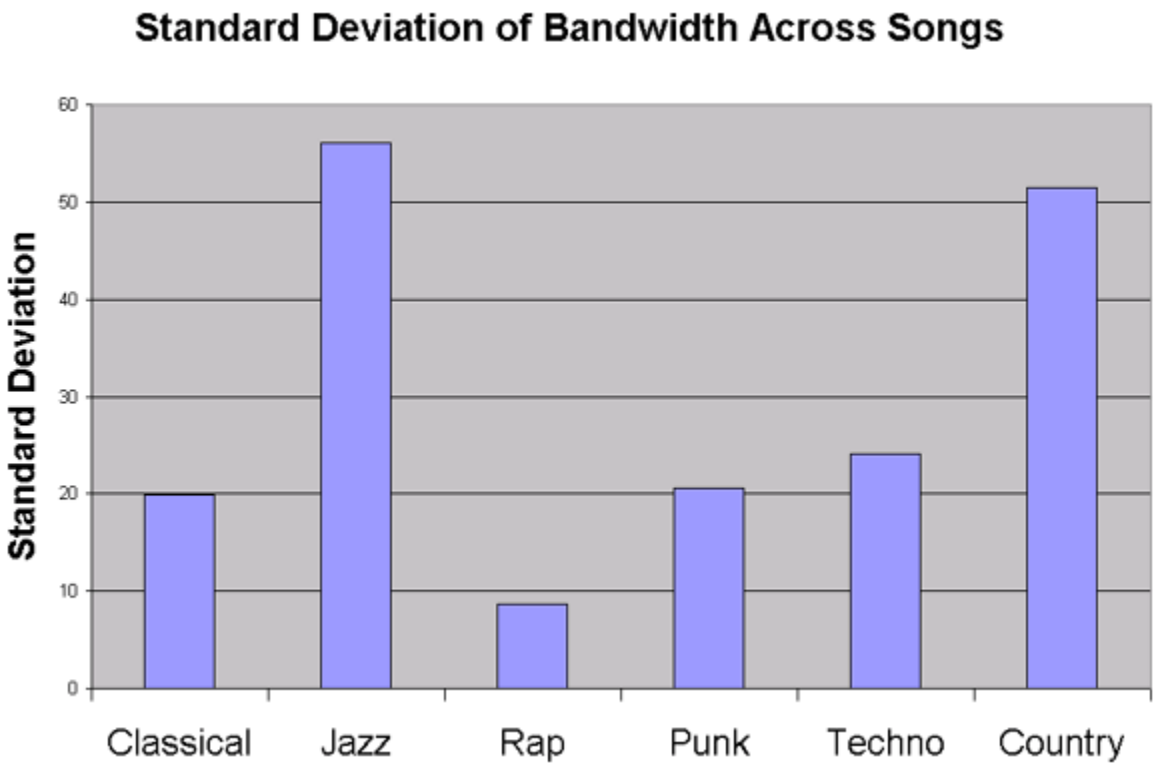
Results

The bandwidth of songs in the frequency domain is very good at distinguishing jazz. However, the bandwidth of punk, techno, and country, all hover around the same value. Rap has most of its power in low frequencies, and those coefficients will be large. Therefore, the bandwidth will be small because the spread of the frequency components is localized in the low frequencies. Jazz, however, has high and low frequencies, so there could be a frequency component in the high frequencies that is large, increasing the bandwidth.

We also give the neural network a measure of how bandwidth changes over the time period of a song. The standard deviations are good at telling jazz and country apart from the other genres, but no one genre stands out.



Overall, bandwidth is a good detector for jazz and rap, but poorer in distinguishing between classical, punk, techno, and country, which all have about the same bandwidth.



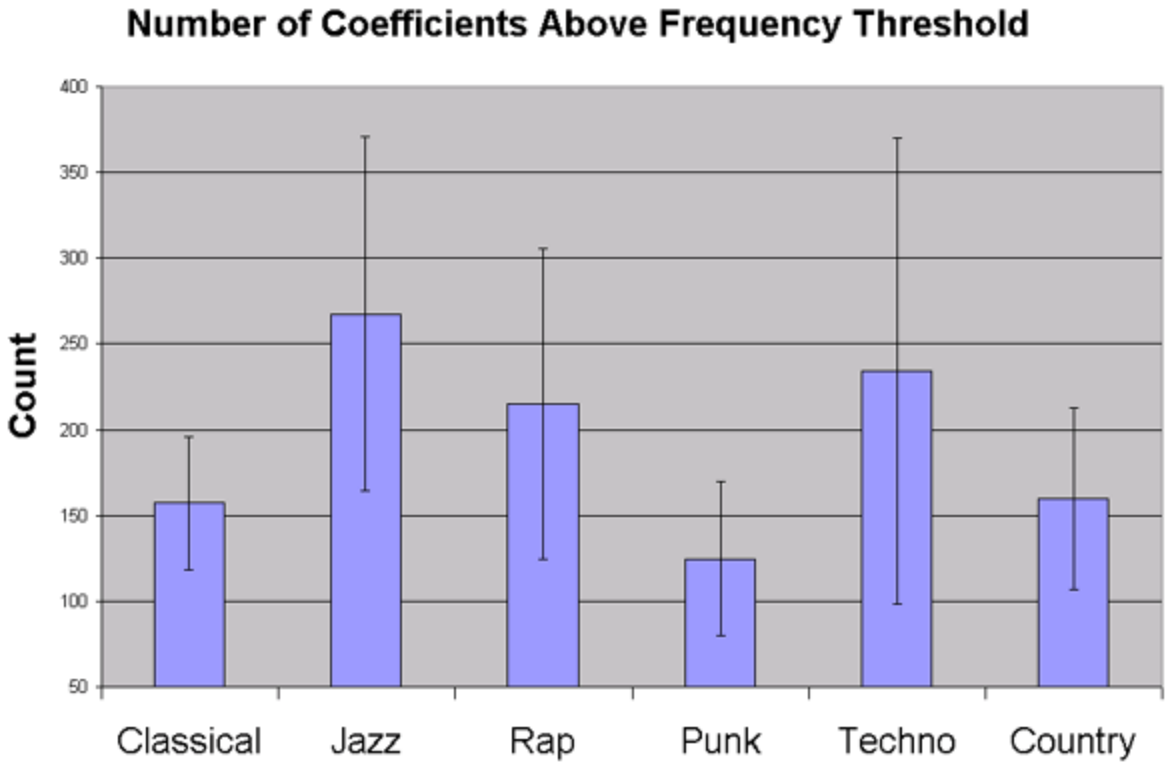
Variation of bandwidth across time.

Music Classification by Genre: Frequency Cutoff

Like most of our other filters, the frequency cutoff had mixed results that varied on the genre in question. Some samples it is readily able to identify, while others if finds quite difficult to pin point directly. For instance, if you fed the filter a sample of classical and a sample of techno, it would have no problem telling you the difference between them. This is because techno has a majority of its energy concentrated at only a few frequencies while classical has its power spread more evenly over a wider band. On the other hand if you were to input samples of punk and country, the filter might tell you that The Ramones sound like Hank Williams. Looking at these results though is not the whole story. A more telling relationship is isolated when the Standard Deviations of these outputs are analyzed. It becomes difficult to isolate any one genre but it does separate them into two main categories:

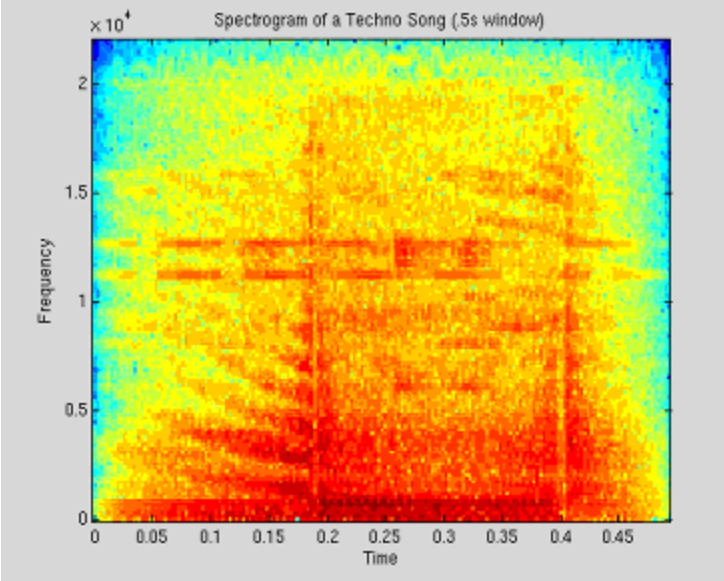
1. Classical, Punk and Country
2. Techno, Jazz, and Rap

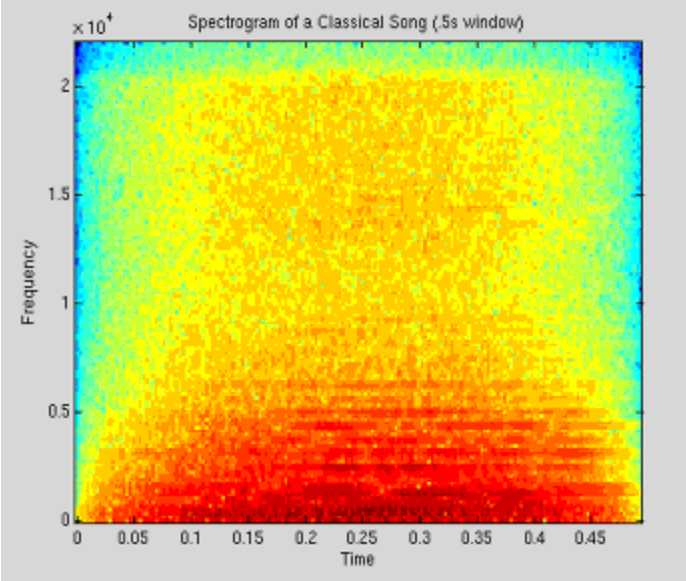
Group one consists of the genres who retained only 40-50 coefficients above the thresh hold, while the genres of group two consistently preserved at least 90 coefficients per sample. This wide gap between them should paint a fairly clear picture of the differences between genres with respect to their cutoff frequencies. This alone isn’t very helpful, but when used in conjunction with other filters, this could prove to the first step in a very powerful tool to help classify music.


Number of Coefficients Above Frequency Threshold
Count
400
350
300
250
200
150
100
50
Classical
Jazz
Rap
Punk
Techno
Country

Music Classification by Genre: Frequency Smoothness

A spectrogram is a tool that belongs to a set of tools called time-frequency representations. Music, on a CD, is a time-vector. Performing an FFT of this time-vector would give us its frequency content. However, a single FFT would lose all time information since it gives us the frequency content of the time-vector as a whole. We need something like an instantaneous frequency response so we have both frequency and time information. A spectrogram essentially breaks a signal up into many different time-vectors and performs FFTs of each. These FFTs are then placed as columns in the spectrogram. In the end, we have a time-frequency representation of our music.

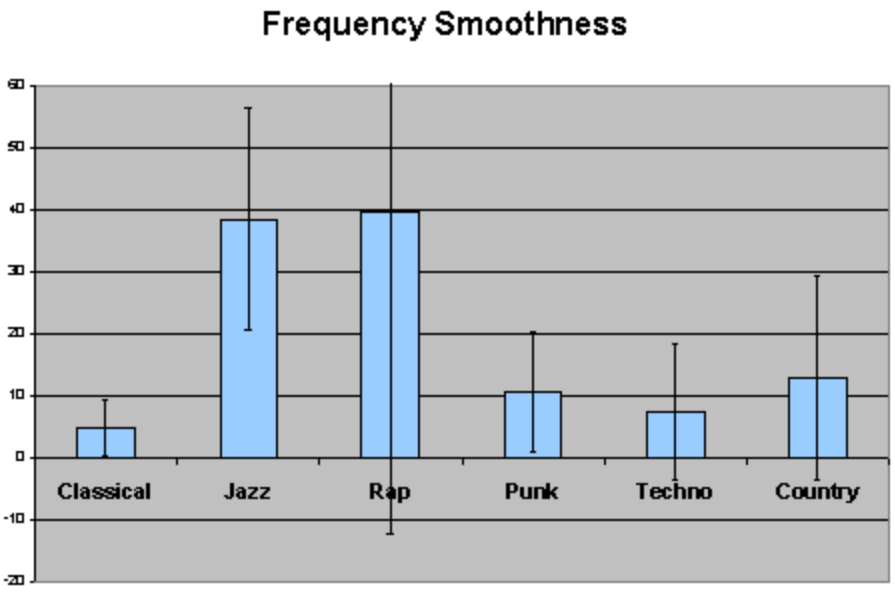




This is a spectrogram of a techno song and a classical song. freqsmooth.m quantifies the differences seen in these spectrograms. To do this, freqsmooth calculates the variance in the indices of the max values of each column. In other words, a song with a clear, loud melody will show small variance in these indices while a song with a harder-to-identify melody will show a large variance.

Results

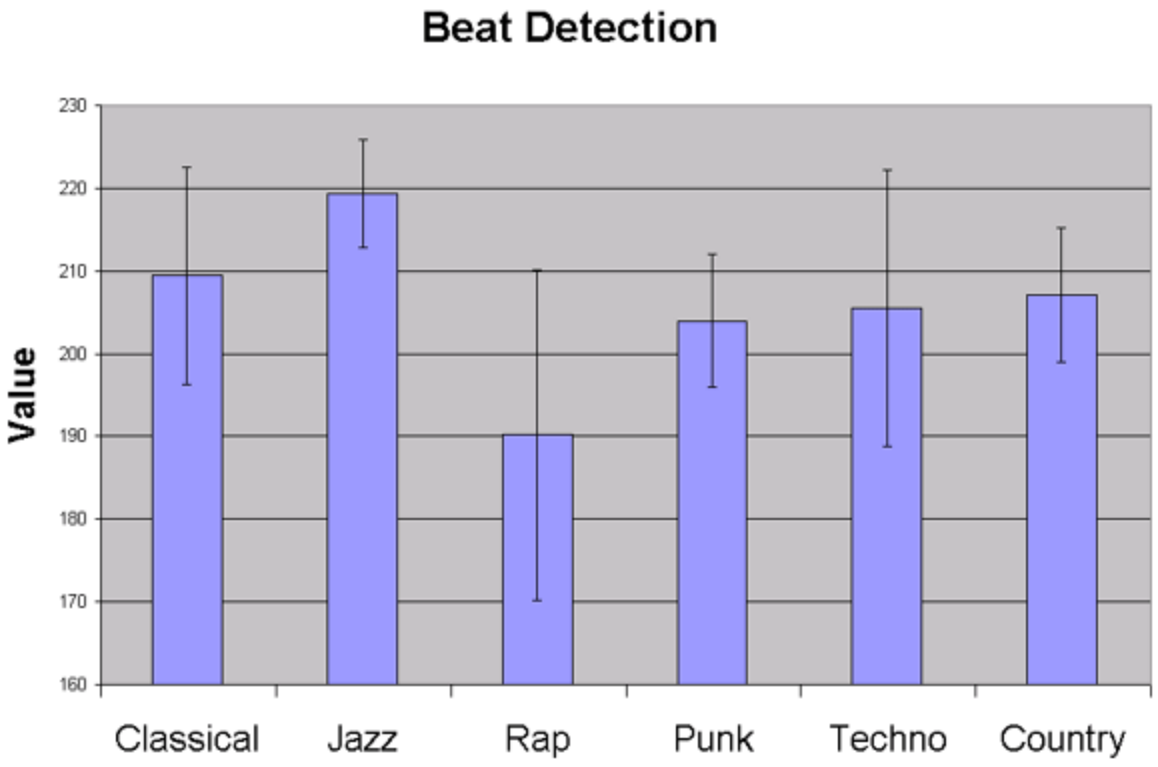
While freqsmooth does give a different value for each genre, it also gives a radically different value for songs within a given genre. In other words, it does not give a good representation of a genre as a whole. Given the plus and minus standard deviation bars, each genre overlaps heavily.

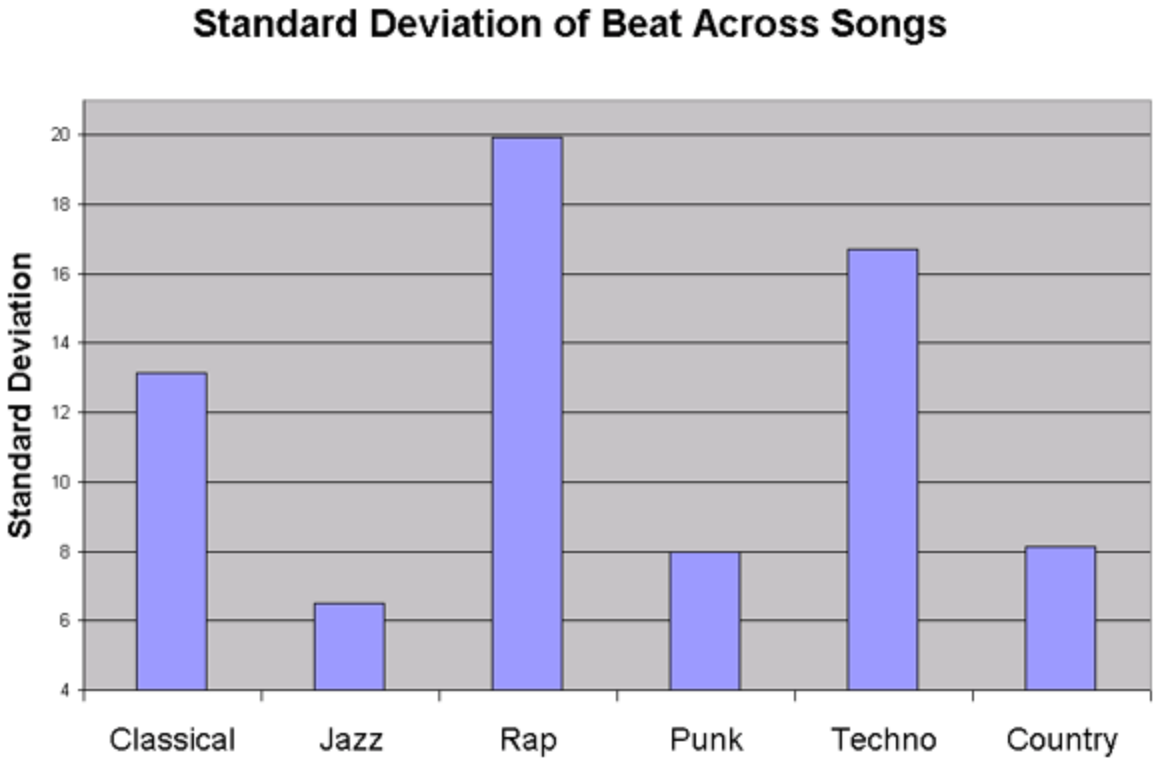

Frequency Smoothness
60
50
40
30
20
10
0
-10
-20
Classical
Jazz
Rap
Punk
Techno
Country

Music Classification by Genre: Beat Detection

Beat detection emphasizes the sudden impulses of sound in the song and then finds the fundamental period at which these impulses appear. It convolves a signal with itself and finds frequency peaks. Then it measures the distance between these frequency peaks. This is done by breaking the signal into frequency bands, extracting the envelope of these frequency-banded signals, differentiating them to emphasize sudden changes in sound, and running the signals through a filter to choose the highest energy result as the tempo. Variation in tempo, found by detecting beat in different windows of the song, helps determine musical genres.

The filter can only separate rap from all other genres effectively, because it has the steadiest backbeat, consistent across the genre! Classical and jazz have too much variability, which makes sense, considering that each piece is often long and divided into sections.



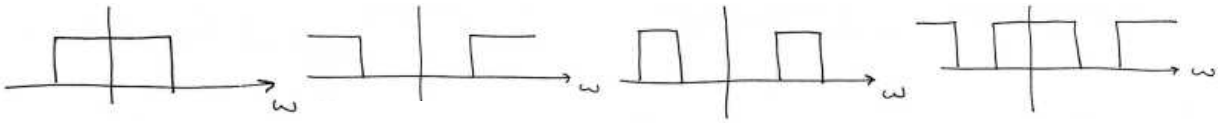

Standard Deviation of Beat Across Songs
Standard Deviation
20
18
16
14
12
10
8
6
4
Classical
Jazz
Rap
Punk
Techno
Country

Ideal Filters

There are four fundamental filters. They are

- Lowpass -- blocks high frequencies, allowing low frequencies through
- Highpass -- blocks low frequencies, allowing high frequencies through
- Bandpass -- blocks all frequencies except those within a certain range
- Bandstop -- blocks only the frequencies within a certain range, allowing all others to pass through

Lowpass Highpass Bandpass Bandstop

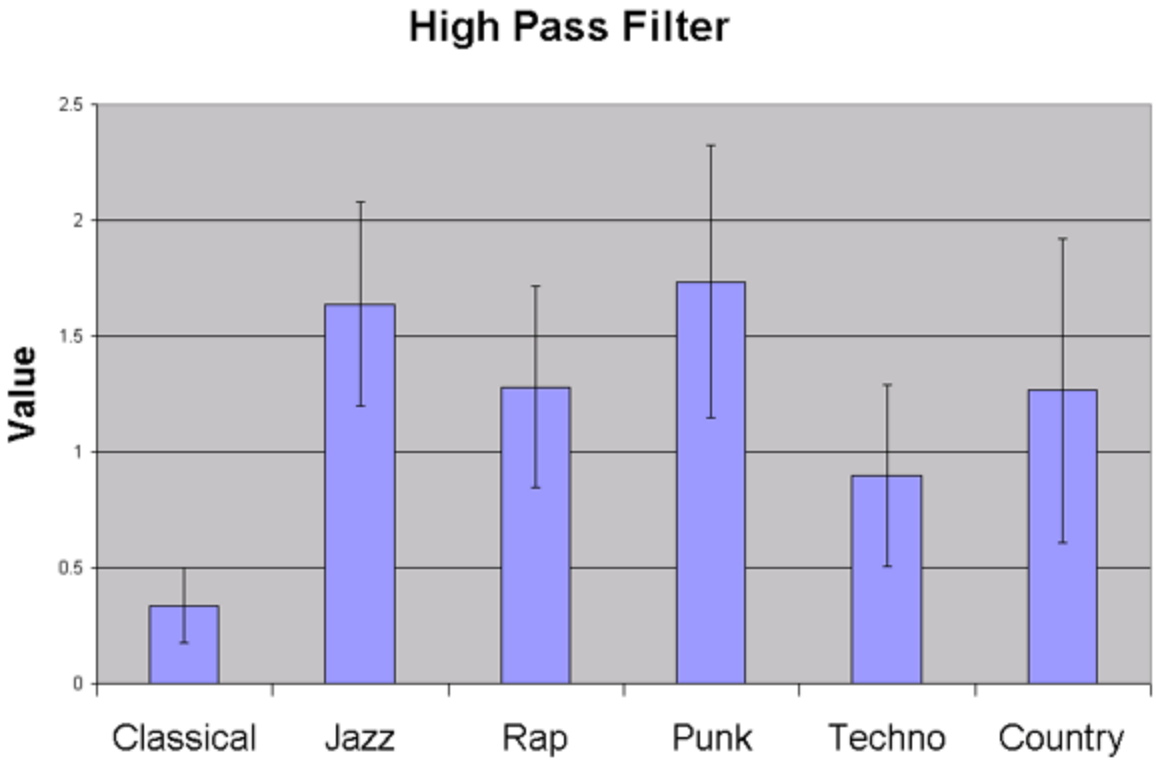


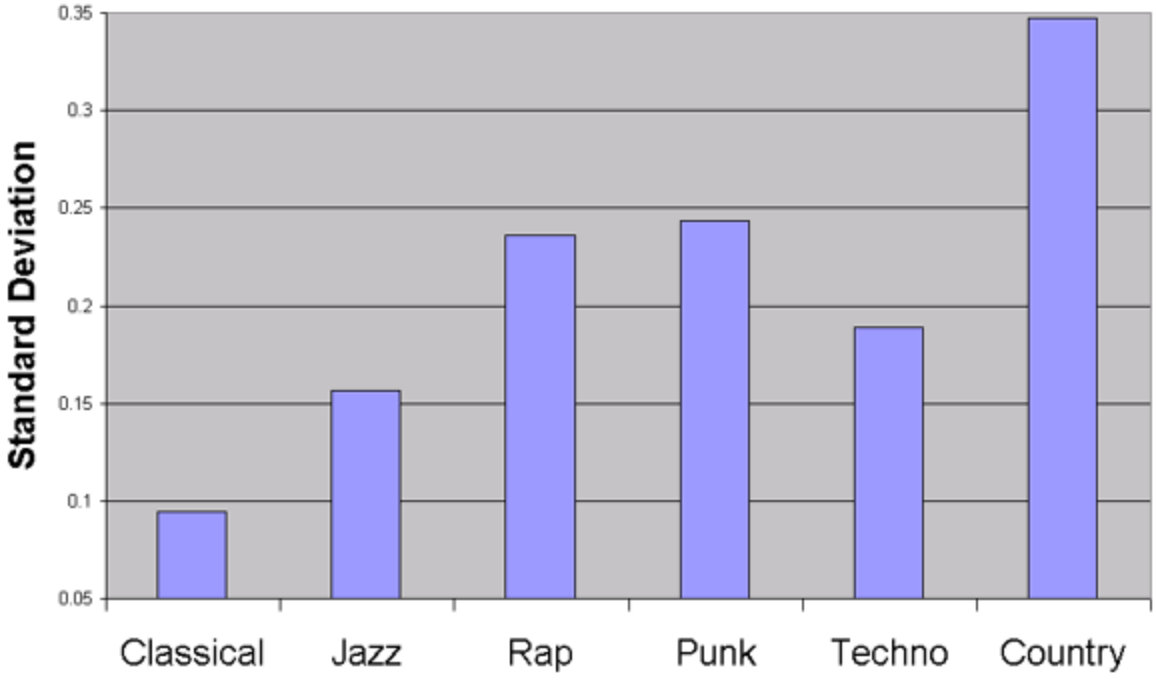
Ideal frequency domain representations of the four fundamental filters.

Another term one may come across in the study of filters is an "allpass" filter. This is one that allows all frequencies through. The only meaningful effect an allpass filter can have is on the phase of the signal.

Music Classification by Genre: High Pass Filter

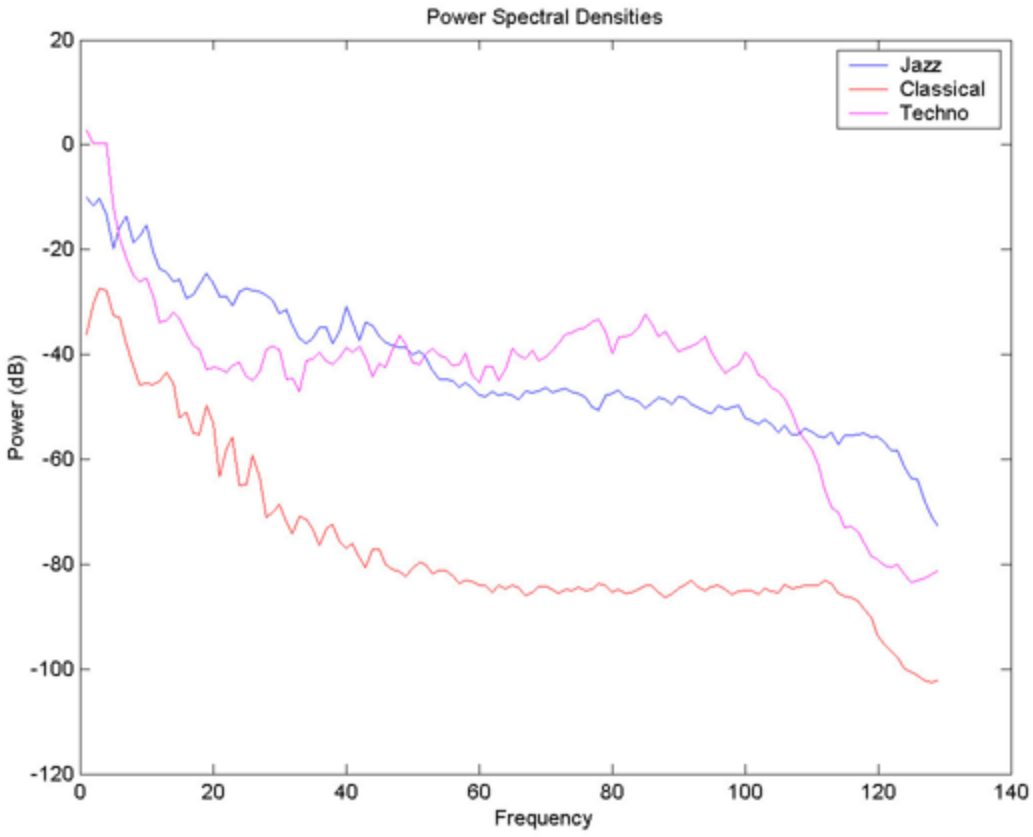
Like most of the filters run on our music samples, the High Pass filter does a good job of identifying some genres, while it has difficulty with others. Like one would expect, classical had the smallest error of any genre tested. This should be intuitive since it uses the lower frequency part of the spectrum. One can think of classical music as being very fluid with few sudden changes in frequency. Conversely, punk and jazz had the highest amount of error, which is a good indication of higher frequencies being utilized. Compared to classical music, these genres are much less fluid and often exhibit rapid changes in tempo. Somewhere between these two extremes are techno, rap and country. The filter has an especially tough time telling the difference between the latter two. Who would have ever thought Garth Brooks and Tupac might get confused with one another. Overall, while the filter cannot explicitly identify the different genres, it does give the user a starting point to isolate between two main groups. This means that another tool must be used along side the high pass filter in order to obtain an effective music matcher.




Standard Deviation of High Pass Filter over Songs
Standard Deviation
0.35
0.3
0.25
0.2
0.15
0.1
0.05
Classical
Jazz
Rap
Punk
Techno
Country

Music Classification by Genre: Power Spectral Density

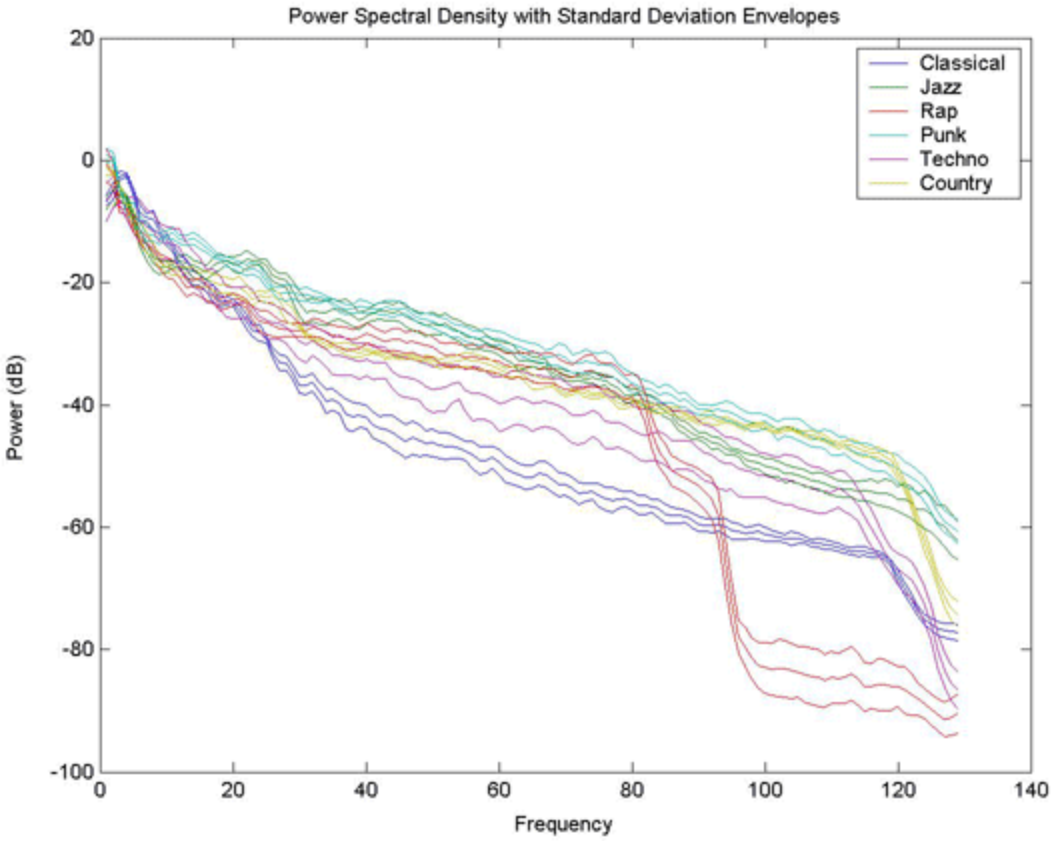
Our program essentially breaks the time-domain signal into windows and computes the norm squared of the FFT of each window. It then averages the magnitude squared of the FFT coefficients of each window, then represents it in decibels. We then have a vector approximately length 100 that represents the power in the frequency domain. This is a measure of exactly what frequencies are present and at what magnitude. Rather than using a single number to characterize the whole signal, our power spectral density program returns a vector representing more subtle changes in the spectrum. The decibel scale helps distinguish and differentiate between genres even further, fanning out the differences between genres.



Results

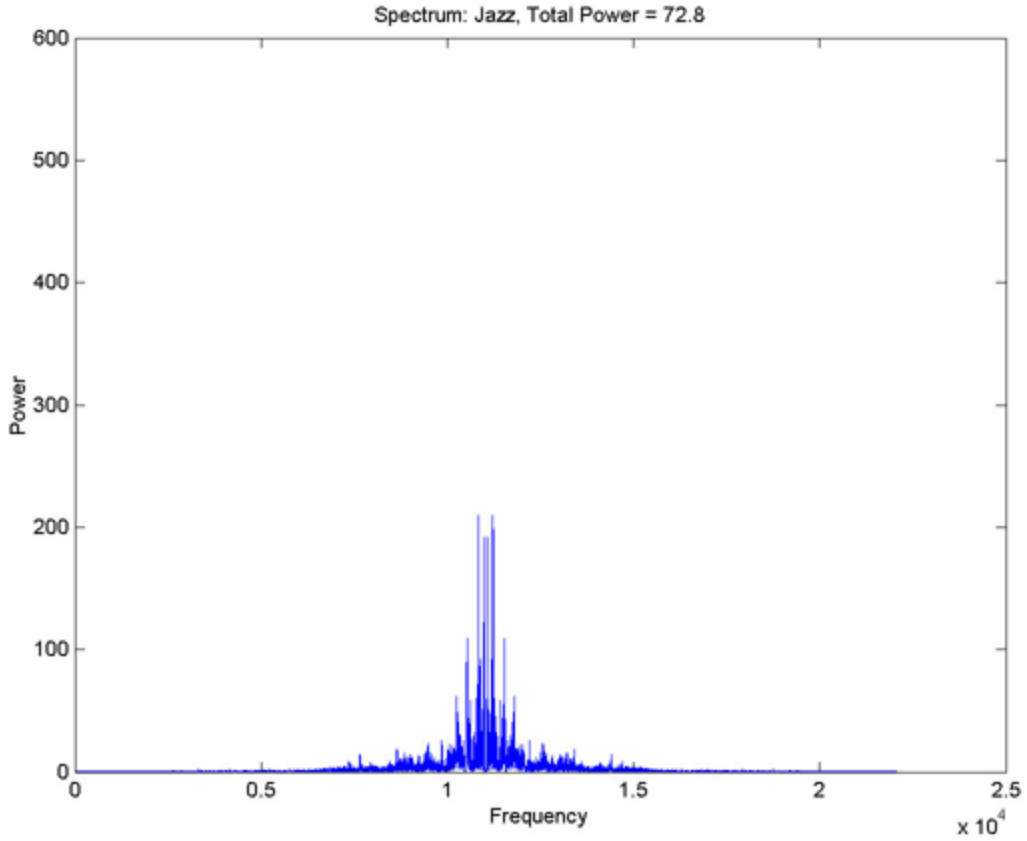
The power spectral density was great at showing patterns between genres. Rap has the most distinct pattern, with a sudden downward slope (red).

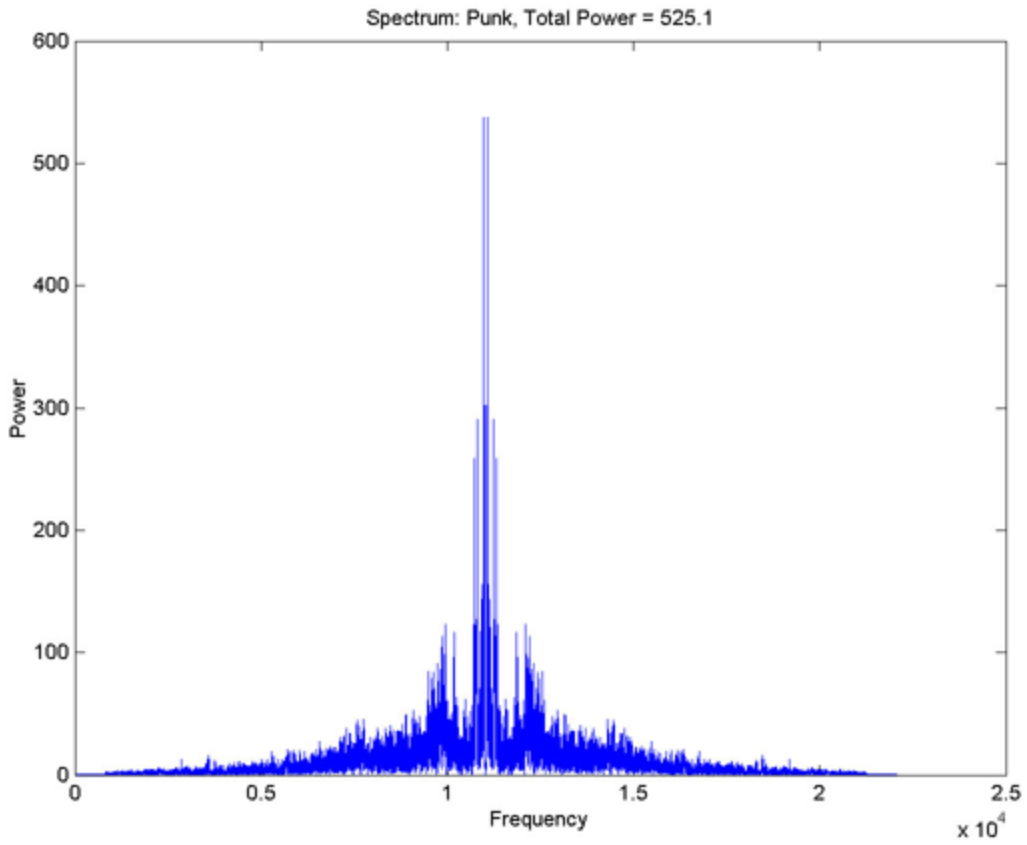
Classical also had a distinctive pattern, with the smallest power at all frequencies. Jazz, punk, and country are all near each other, but at higher frequencies, begin to fan out. Looking closely at the envelopes, techno spans the largest area, encapsulating almost all of jazz, punk, and country. This is one reason why techno could not be distinguished very well from those genres.



Music Classification by Genre: Total Power

The power in a signal is the norm squared of the frequency components of the signal. The vectors are first normalized to the maximum value in the vector such that we are not analyzing loudness, but more accurately, the L-2 norm divided by the L-infinity norm. It measures how many harmonics are present in the signal and how much of each harmonic. In our case, the music samples have a wide range of total power: classical piano has low power with few harmonics, whereas punk has high power. You can see from these two plots of the spectrum, on the same scale, that jazz has much smaller power. Jazz has fewer frequency components of smaller power.

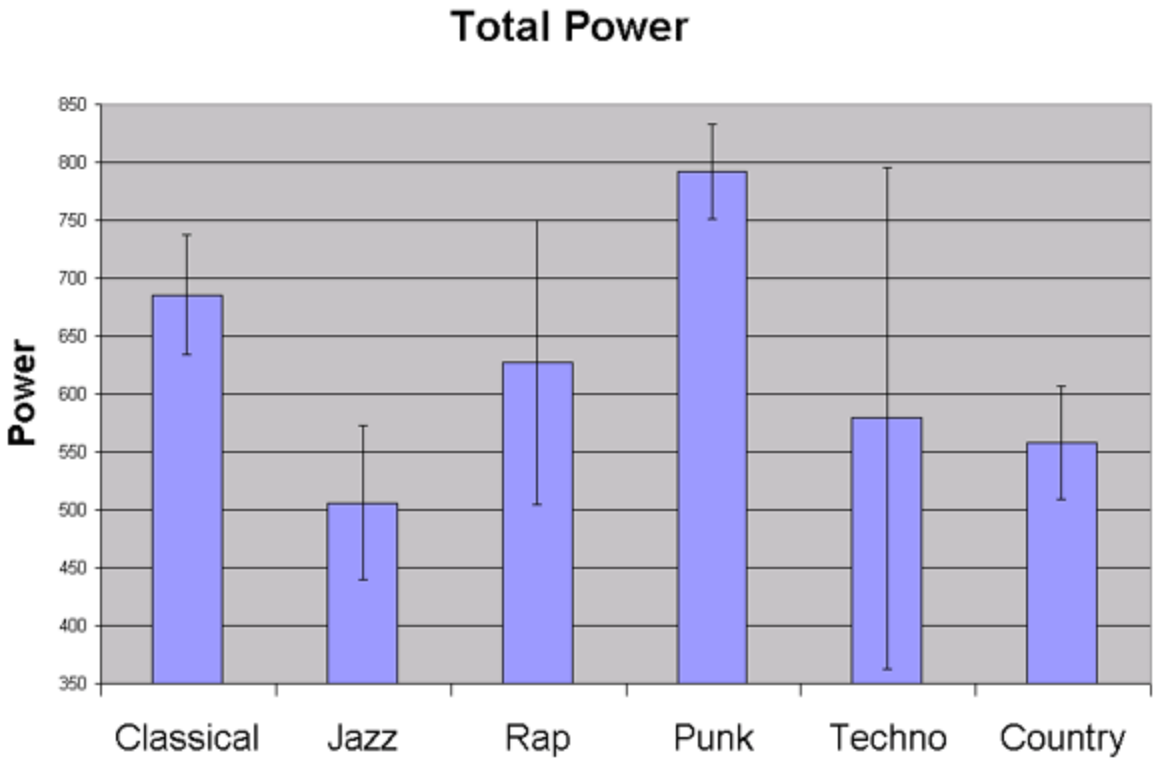




Results

The total power of a signal changes radically between genres. Jazz has the lowest total power, while punk tops the list. Punk also has the lowest standard deviation; there should be very little confusion with the rest of the genres. Techno is the least discernable: its standard deviation encapsulates all the other genres. Looking closely at the graph, the spread of the standard deviations of classical and country does not encapsulate any other genres, so they should be easily identified.

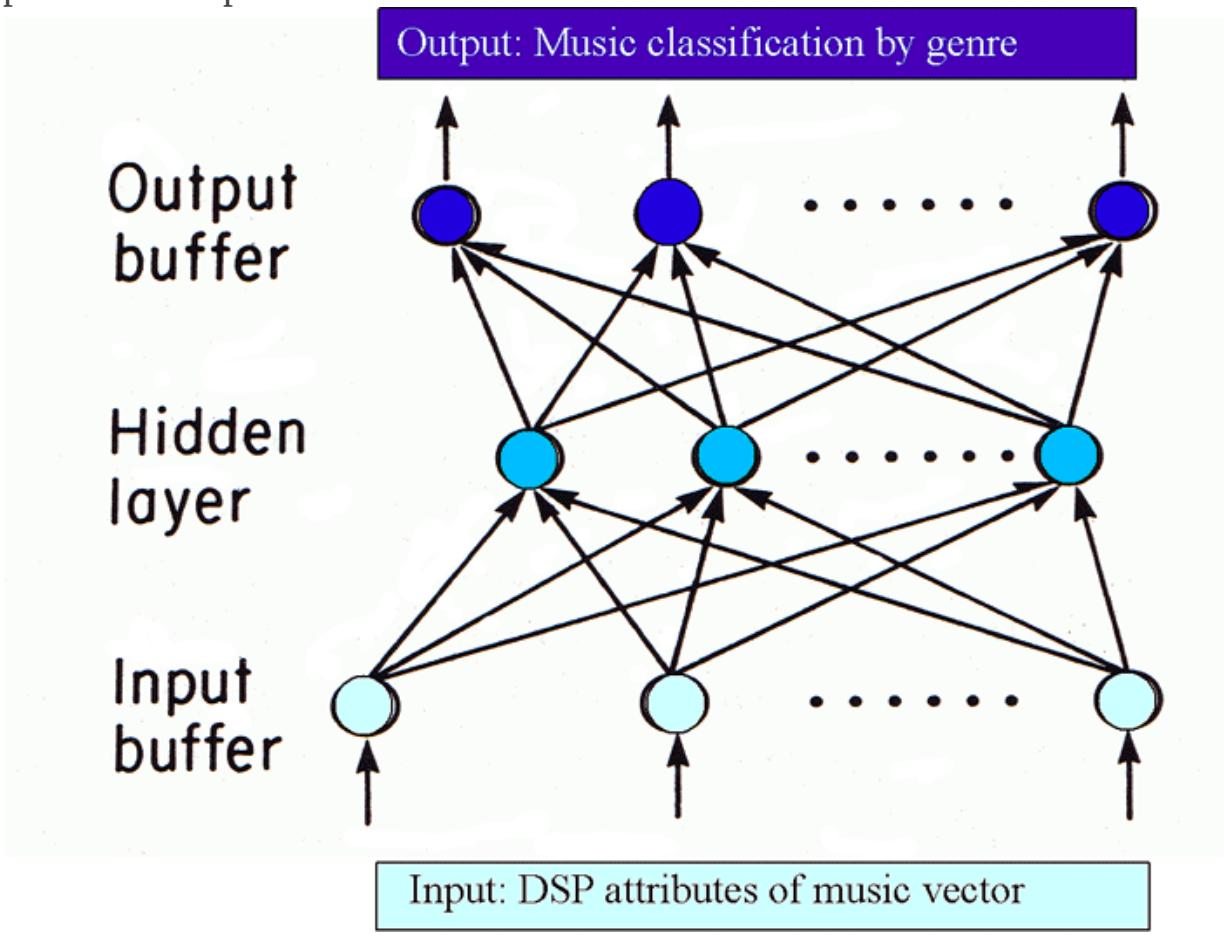
The standard deviations of rap and techno are very distinct, whereas the others are all about the same value. Although the average total power of techno may not be a good indicator, the standard deviation should be able to pick out techno.


Total Power
Power
850
800
750
700
650
600
550
500
450
400
350
Classical
Jazz
Rap
Punk
Techno
Country

Neural Networks

At their core, neural networks are pattern recognition systems. They predict an output given a sequence of inputs and their corresponding classification. They are based on biological nervous systems, in which there are many inputs and numerous outputs to a single neuron. On the highest level, the neural network is a primitive learning machine that can be used to process data such as stock market quotes, DNA sequences, and in our case, music classification.

Neural networks are systems that take a lengthy input, process the data, and predict an output.



The processing is done by multiple, weighted layers of nodes. Each node is connected to every node in the next layer, and at each interface between nodes are connecting fibers weighted by a sum. Neural networks are given a vector of inputs, usually longer than the output. The first layer of nodes is

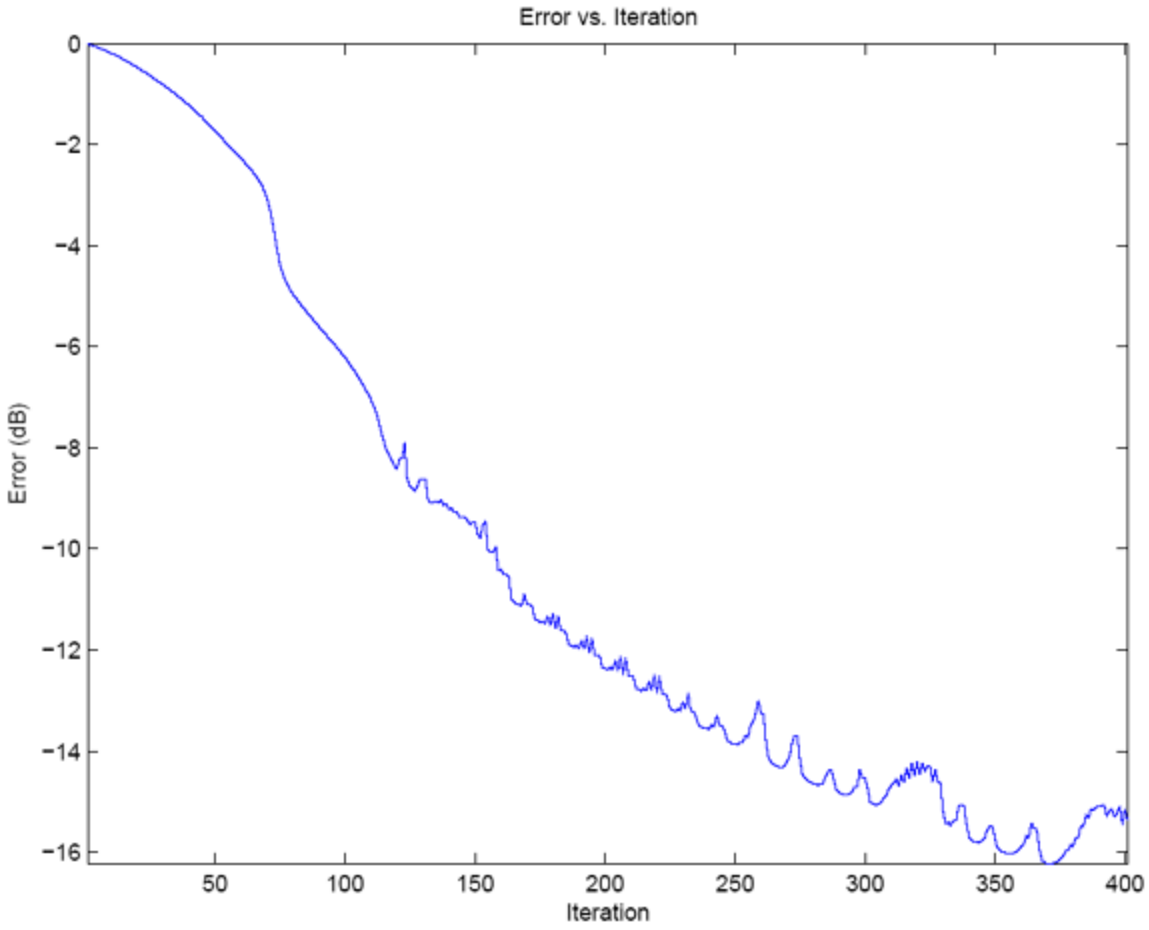
the same length as the input. The nodes at each successive layer sum their inputs, weight the sum, and produce an output. The output of the final layer is the output of the system. In this manner, an output is predicted given an input.

The remaining question is how the weights are determined. The use of neural networks is twofold: you must first "train" the network by giving it inputs and their corresponding outputs, and then you may test the network by giving it inputs with no outputs. The training determines the weighting on the nodes. For example, we train the neural network by giving it the vectors of signal processing data (bandwidth, power spectral density, etc.) as well as the corresponding classification of music. Classical music is denoted as [1 0 0 0 0 0], jazz is denoted as [0 1 0 0 0 0], etc., as shifted delta functions.

There are many methods to train neural networks, but the one we use is called backpropogation. The neural network takes the input and feeds it through the system, evaluating the output. It then changes the weights in order to get a more accurate output. It continues to run the inputs through the network multiple times until the error between its output and the output you gave it is below a defined tolerance level.

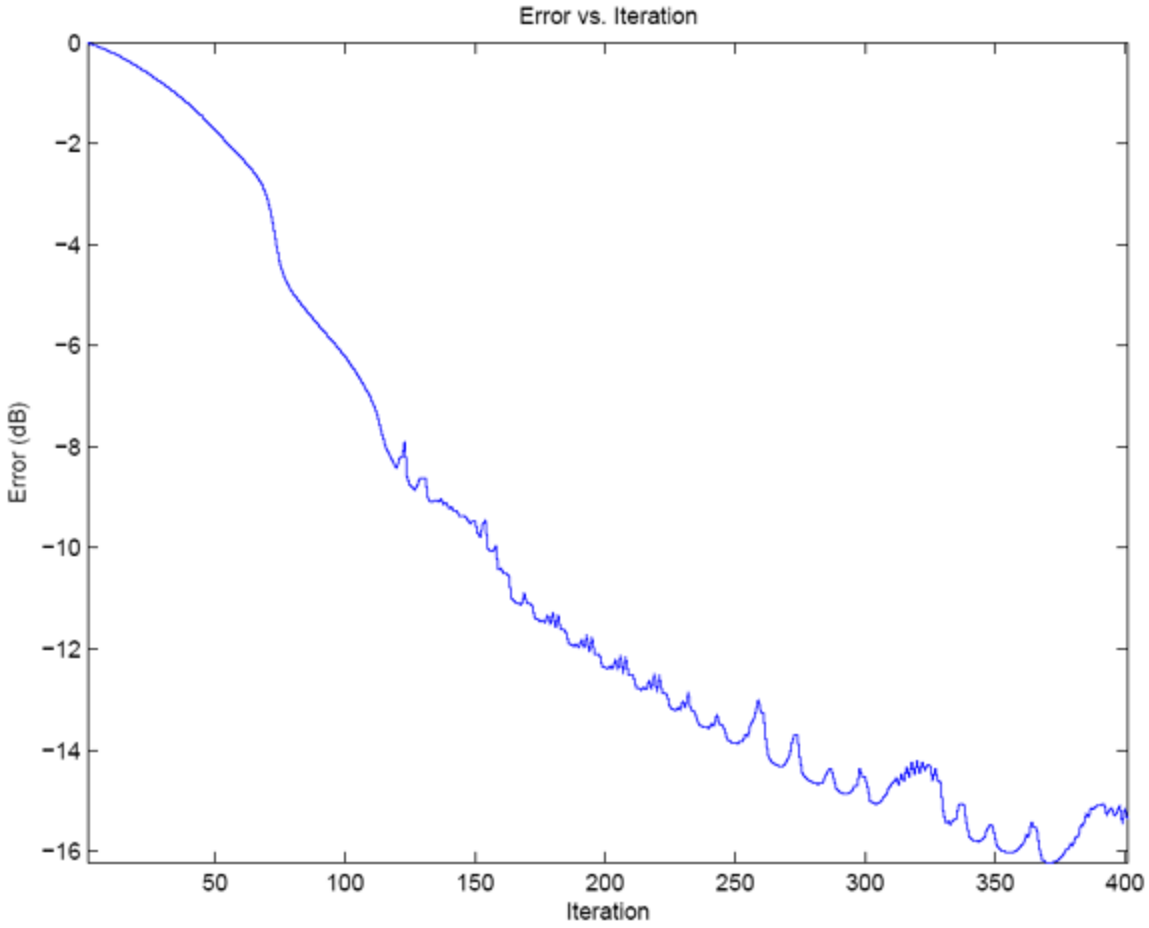
After training is completed, we test the neural network with songs that it has never seen before. It predicts a classification of genre based on the weights it created during training.

To train the neural network, we used the method of back propagation. There were 15 nodes in the hidden layer, and we used an adaptive learning rate training function. This means that the network analyzed its learning rate after each iteration, changing it to remain relatively constant. For instance, if the learning curve is too steep, and the network is learning too quickly, it decreases its learning rate, and vice-versa. This is a graph of the error (learning rate) versus time:

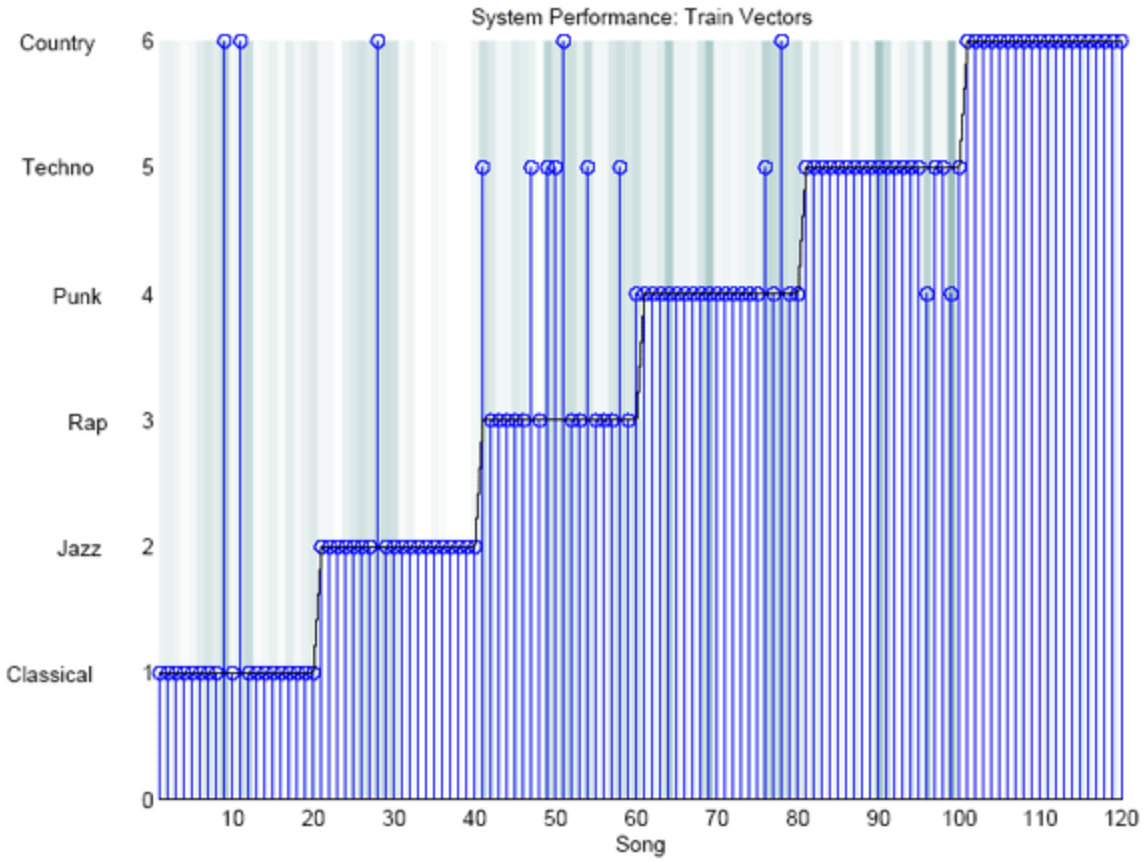


Error in the neural network decreases with each successive iteration.

Music Classification by Genre: System Performance

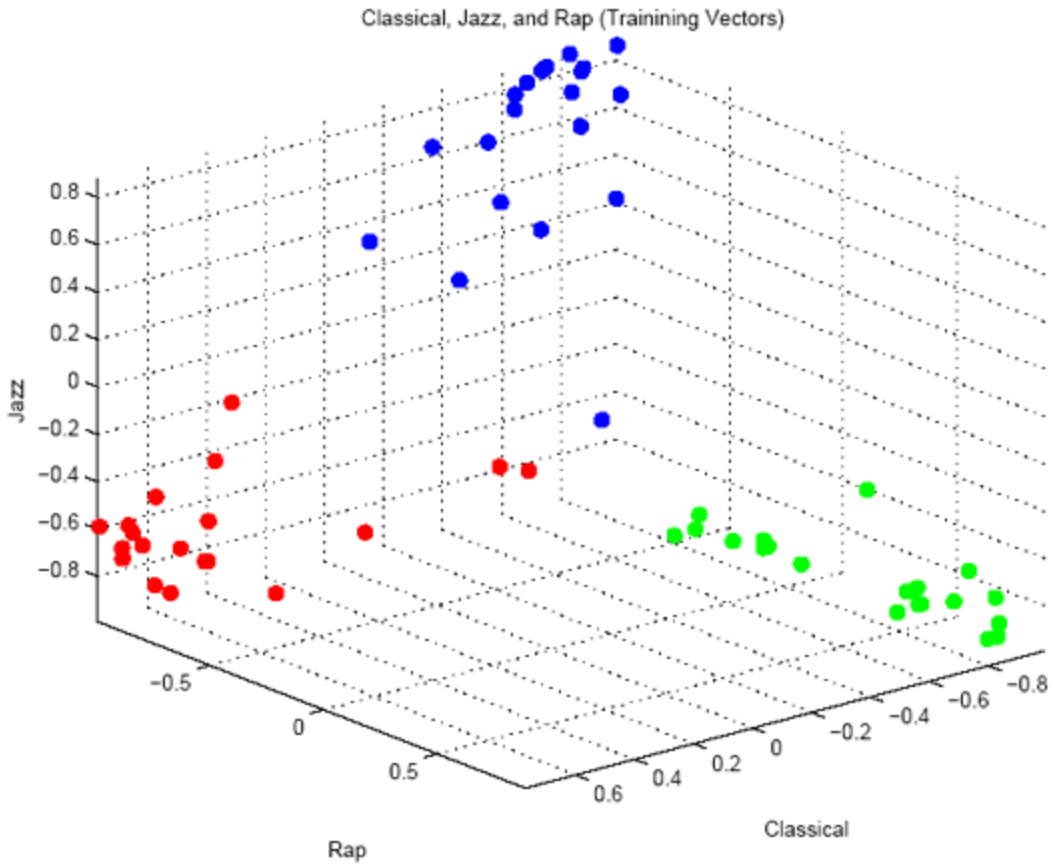


Performance of the neural network improved with successive inputs of training vectors. It begins to recognize characteristics of each music genre!

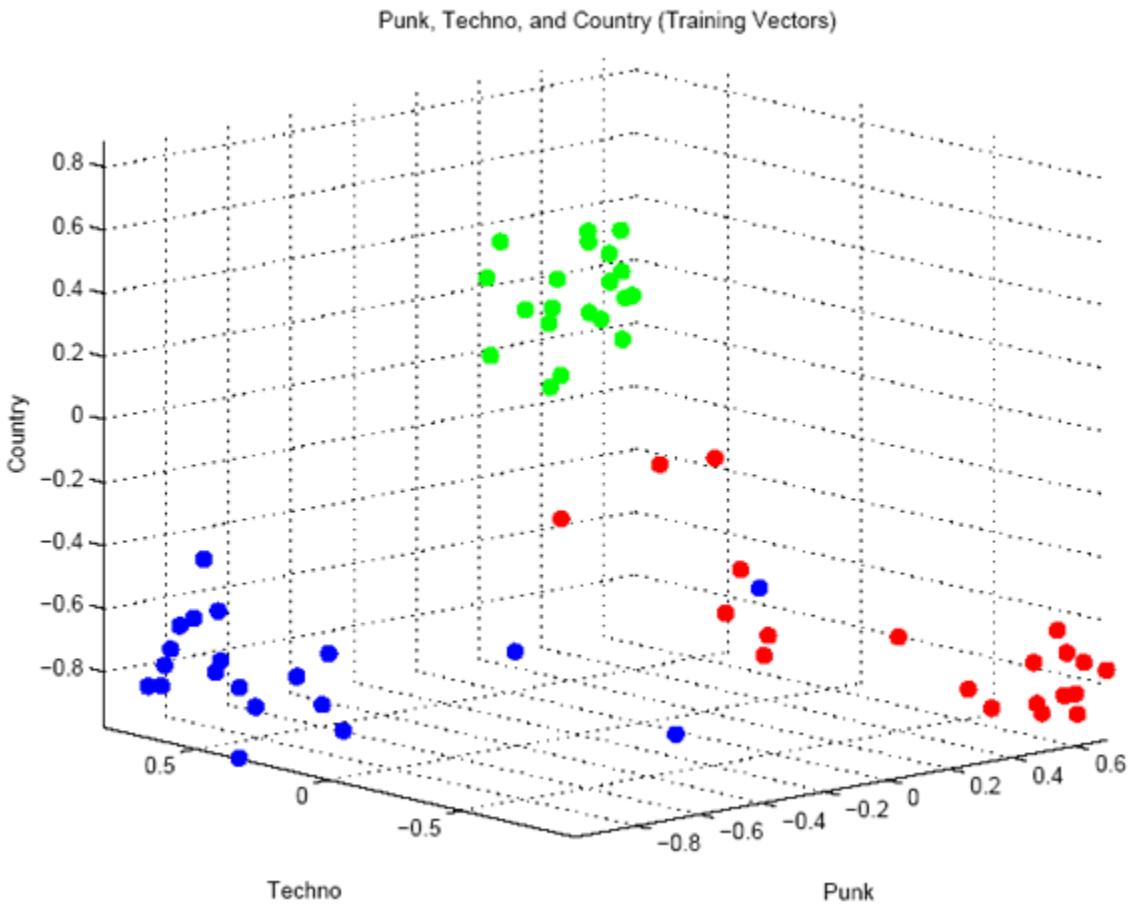


When tested with the training vectors, the system is 87.5% accurate. Higher accuracy implies that the system has memorized the training set and is unable to generalize when given new inputs. Lighter background stripes indicate greater certainty in identification, while increasingly darker hues note greater uncertainty. Horizontal black bars indicate actual genre, and the stems indicate predicted genre.

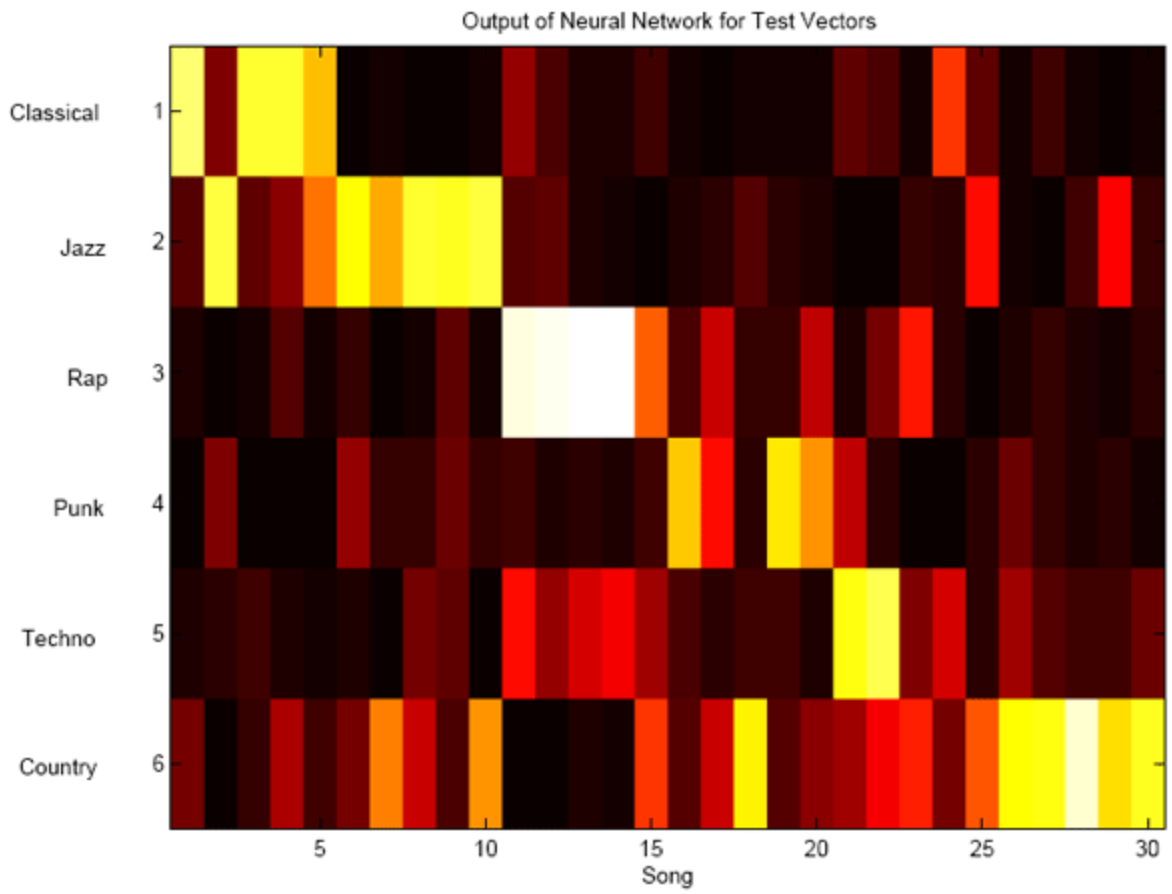
These plots show how well the first three genres separate in the output of the network. Even the testing vectors are separated by a high degree of confidence.



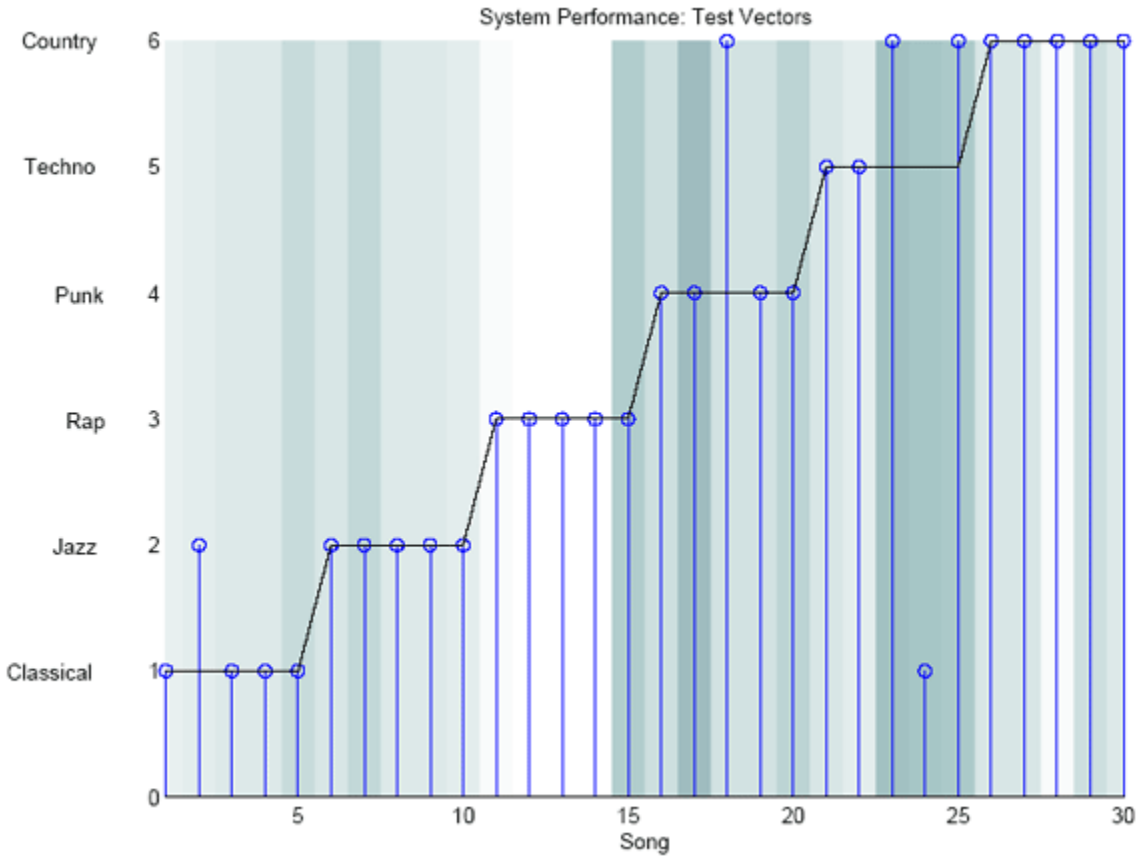
Spatial separation, weighted by sureness level, of classical (red), jazz (blue), and rap (green) in the training vectors.



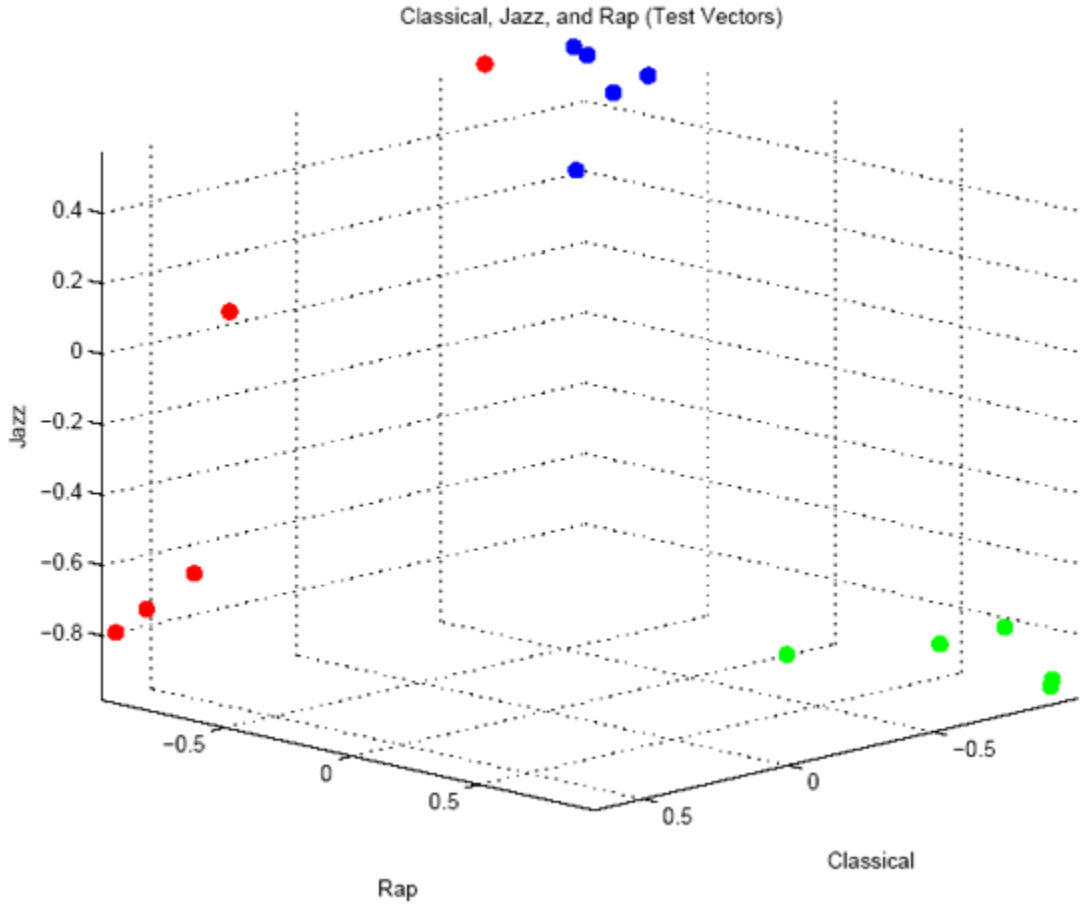
Spatial separation, weighted by sureness level, of punk (red), techno (blue), and country (green) in the training vectors.



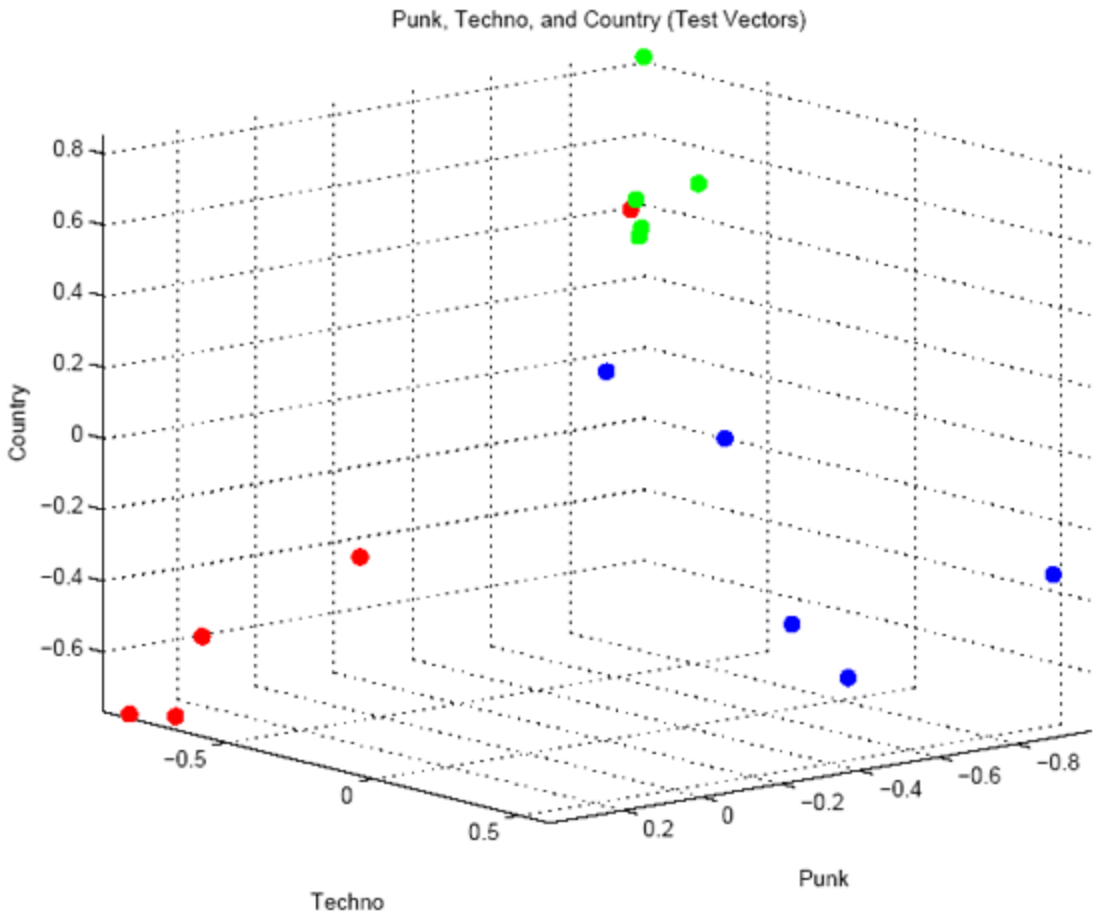
Though worthy of the Museum of Modern Art, this depicts the output of the neural network for each of the six genres.



Provided songs that the network has never seen, it performs perfectly and with high confidence for rap, while classification of techno is comparably poor. However, the system is aware of this: error coincides with lack of confidence. Lighter background stripes indicate greater certainty in identification, while increasingly darker hues note greater uncertainty. Horizontal black bars indicate actual genre, and the stems indicate predicted genre.



Spatial separation, weighted by sureness level, of classical (red), jazz (blue), and rap (green) in the training vectors.



Spatial separation, weighted by sureness level, of punk (red), techno (blue), and country (green) in the training vectors.

Our system successfully determined the genre of the vast majority of the test songs. Not only did the system choose a genre, it quantified its output with a level of sureness. When the system was in error, there was a corresponding uncertainty. The genre that gave our system the most difficulty was techno. The output of the DSP functions for techno had a very high standard deviation, making it hard for the neural network to distinguish its pattern from those of the other genres. Given an unknown song, there is an 84% chance that the system can determine the genre successfully.

Back propagation mathematics

Error definition

The back propagation method is the example of the wide class of training methods based on the information covered in the gradient of error function. The independent variables in this minimization are weights of neural network and the considered error to be minimized is the root mean square one.

Let us consider the training set composed of L ordered pairs, of the following form: $\left\{\left(\mathbf{x}^{(1)}, \mathbf{d}^{(1)}\right), \left(\mathbf{x}^{(2)}, \mathbf{d}^{(2)}\right), ..., \left(\mathbf{x}^{(L)}, \mathbf{d}^{(L)}\right)\right\}$ Furthermore, let us define the total error E generated on outputs of neural network after presenting the entire training set, as:

$$E = \sum_{l=1}^{L} E^{(l)}$$

where:

$$E^{(l)} = \sum_{m=1}^{M} E_m^{(l)} = \frac{1}{2} \sum_{m=1}^{M} \left(d_m^{(l)} - y_m^{(l)}\right)^2$$

As was already told, the independent variables in the minimization of error E are weights w_{ij} Since even for the relatively small networks the number of weigths is big, in real applications, the training of the neural network is the minimization of the scalar field over the vector space with hundreds or (more often) thousands dimensions. One of the minizmiazation techniques for such problem is the steapest descent method

$$\int_0^1 x^2 \, \mathrm{d}\, x$$

$$\sum_{n=1}^{\infty} 2 \times 2^{1/2} \left(26390n + 1103\right) (4n)! \Big(9801 \times 396^{4n} (n!)^4\Big)^{-1}$$

$$\sum_{n=1}^{\infty} \frac{2\sqrt{2}\,(26390n+1103)\,(4n)!}{9801396^{4n}n!^4}$$

$$\sum_{n=1}^{\infty} \frac{2\sqrt{2}\,(26390n+1103)\,(4n)!}{9801396^{4n}n!^4}$$

$$\sum_{n=1}^{\infty} 2 \times 2^{1/2}\,(26390n+1103)\,(4n)!\Big(9801 \times 396^{4n}(n!)^4\Big)^{-1}$$

Chris Hunter

Chris in a drysuit.

Hi everyone, I am an Electrical Engineering major and a member of the Class of 2006 at Rice University. This bio is an informal section about my interests and my life. If you would like a resume, please feel free to email me at chunter@rice.edu

I was born in Tampa Bay, FL, but I moved to Austin, TX quick enough such that I don't even remember Florida. Austin, in my humble opinion, is one of the most gorgeous places in the country. The weather is great and the people are friendly. If you ever feel like visiting the city, email me and I can show you around.

I am a person of many interests. Let's see, I have had classical piano training for about 11 or 12 years. While I do not continue formal training, I

have branched out to music composition. If you would like to hear a few of my works, stop by http://www.mp3.com/chrishunter. On the more active side of my hobbies, I enjoy power-kiting (large pull-you-off-the-ground kites) and wakeskating/wakeboarding. The latter is my most recent passion, but I'm off to a quick start. In case you are curious, wakeskating is like skateboarding on the water (i.e. you are not bound to the board like you are on a wakeboard). Thankfully, Austin is one of the top places for water sports thanks to our ample natural lakes and mild climate. I am by no means a great wakeskater or wakeboarder, but here are a few pictures of my adventures:

Wakeboarding on Lake Travis.

Cold water requires a wetsuit.

Colder water requires a drysuit.

Let's not talk about it.

Melodie Chu

Aloha! My name is Melodie Chu and I’m a sophomore at Sid. I’m from Honolulu, Hawaii, and live about 10 min from the beach. I don't know how to surf, but I enjoy body boarding. Other than water sports, I like running and photography. Incredibly, I saw snow for the first time last spring and tried my luck at snowboarding. The experience was quite unlike body boarding, but I can’t wait to make another snowman. After graduating from Rice, I plan on getting a master’s degree and working in industry.

Hawaii is a beautiful island. To see amazing views from around the island, visit www.owlnet.rice.edu/~mchu.

Mitali Banerjee

Dressed in green lengha, traditional Indian dress.

Senior at Rice University double majoring in electrical engineering and anthropology, I have interests that extend beyond physics and robotics to include such hobbies as reading, writing, sketching, playing piano, and conversing with interesting people. My passions include chocolate, Steinbeck novels, skeeball, photography, literary trivia, classic movies, black currant jam, and toe socks.

My forty minutes of fame: appearing on College Jeopardy! as a freshman. With the help of Shakespeare and denim miniskirts, I met Alex Trebek while fulfilling one of the items on my things-to-do-before-twenty-one list, a high school Quiz Bowl fanatic's dream!

JIBA!!! Baker might have its elegant commons, Hanszen might have its weenie loft, Lovett might have a view of the medical center, Martel might have the accommodations of a five-star hotel, Sid Rich might have half-level floors and a friendly atmosphere, Wiess might have its team spirit and war pig, and Will Rice might have... wait, what does Will Rice have?... However, Jones Sweet Jones, be it ever so humble, there's no place like Jones. Even though I'm off-campus I still consider it home base for this final year.

Brief academic history: as a chubby star, I went to John F. Ward Elementary, starting third grade the first year it opened, transferring from Clear Lake City Elementary, where I started school in Houston after moving from Long Island. Then I became a cardinal at Space Center Intermediate, which has since been rebuilt in a new location on Saturn Lane. Graduated as a Class of 2000 falcon from Clear Lake High School, I enjoy my time (entirely more than I should) at college.

Jordan Mayo

If I don’t have my head buried in a book or in front of a computer screen, I’m probably doing something for ROTC. Originally from Houston, Texas, I am a junior Elec Major from Wiess College. I plan to graduate in May ’05 and will then “donate” a significant portion of my life to the United States Navy. I enjoy camping, and reading, but most of all I’m a pretty boring person.